SELF-SUPERVISED LEARNING FOR IMPROVED SYNTHETIC APERTURE SONAR TARGET RECOGNITION

BW Sheffield Naval Surface Warfare Center, Panama City, USA

1 INTRODUCTION

The application of computer vision in autonomous underwater vehicles (AUVs) presents unique challenges due to the unpredictable and often harsh conditions of marine environments. Traditional computer vision research, which primarily relies on optical camera imagery, struggles to adapt to underwater settings characterized by poor lighting, sediment suspension, and turbidity. Consequently, acoustic sonar, particularly its variant - synthetic aperture sonar (SAS), has emerged as a preferred choice for underwater imaging. SAS-equipped AUVs can sweep seafloors to generate high-resolution imagery, offering a level of detail superior to other sonar types. However, the high-resolution SAS imagery, while rich in detail, is voluminous and presents a significant challenge for labeling, a crucial step for training deep neural networks (DNNs).

DNNs have gained traction in comparison to classical machine learning methods. This is attributed to their ability to autonomously discover features in data. This eliminates the need for manual feature crafting by subject matter experts. A significant limitation of DNNs is their requirement for large volumes of labeled data and considerable computational power to learn robust features. In the context of SAS, not only is labeled data scarce, but it is also not as readily available as it is for conventional camera imagery.

Recently, self-supervised learning (SSL) has gained popularity propelled by the increasing availability of computational power and data. SSL enables models to learn features in data without the need for labels, presenting a potential solution to the data labeling challenge in SAS.

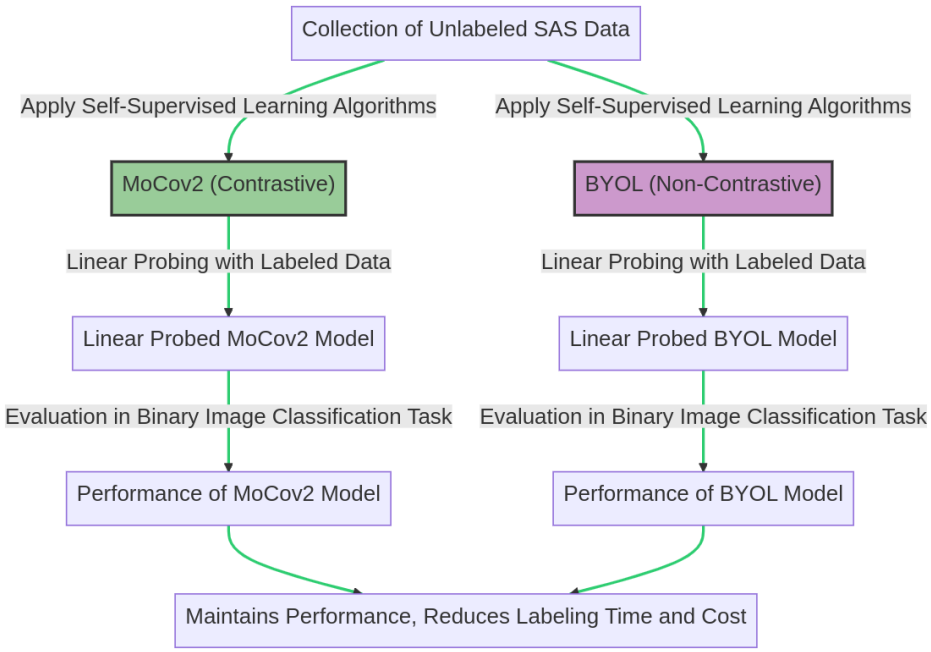

Figure 1: The objective of this paper evaluates two different SSL models in limited label scenarios

This study aims to evaluate the performance of two prominent SSL algorithms, MoCov2 [1] and

BYOL [2], against the well-regarded supervised learning model, ResNet18 [3], for the binary image classification task as shown in Figure 1. The SSL models were trained on real-world SAS data to learn useful feature representations for downstream binary image classification tasks. The findings suggest that while both SSL models can best the performance of a fully supervised model with access to a small amount of labels in limited label scenarios, they do not exceed it when all the labels are used. This study underscores the potential of SSL as a viable alternative to traditional supervised learning, capable of maintaining task performance while reducing the time and costs associated with data labeling.

2 RELATED WORK

SSL has been a burgeoning area of research in recent years, particularly within the remote sensing domain [4–14]. Although SSL applications have significantly advanced across various fields, their application to SAS remains relatively unexplored.

In 2022, Preciado-Grijalva et al. [15] demonstrated the potential of SSL in forward look sonar (FLS) sonar applications can yield classification performance comparable to supervised pre-training in a few-shot transfer learning setup. In the unsupervised and semi-supervised learning domains, researchers have applied methods to reduce the burden of labeled SAS data [16–18].

The potential of SSL in Synthetic Aperture Radar (SAR) applications, a field closely related to SAS, has been demonstrated in several studies [19–27]. These studies have shown that SSL can effectively leverage the vast amounts of unlabeled SAR data to achieve meaningful results.

However, the application of these methodologies remains largely unexplored in the context of SAS data. This gap in the literature may be due to the unique challenges associated with SAS data, such as the sensitive nature of the data and the computational resources required for training.

3 METHODOLOGY

The subsequent experiments are designed to conduct a comparative evaluation of the performance of the models' representations. This is achieved both qualitatively, through the visualization of the latent spaces, and quantitatively, based on the ultimate classification outcomes. To establish a common baseline, all SSL models, MoCov2 and BYOL, use the same ResNet18 backbone.

Training of the pre-existing models was carried out using PyTorch Lightning for up to 100 epochs. This was carried out on eight Nvidia A6000 GPUs, each equipped with 48GB RAM. The DDP strategy helps to improve the consistency of batch normalization across multiple GPUs. Distributed sampling ensures that each GPU processes a unique subset of the total data in each epoch, leading to more stable training and potentially better performance. Unlike other deep learning methods, high batch sizes are highly desirable in achieving good results as the task of the loss function is to pull positive instances together and push negative instances away.

The downstream task for comparison was binary image classification using binary cross entropy for the loss function. A threshold of 50% was used to make a decision on whether an image contained a object of interest. In an iterative manner, the SSL pre-trained model is fine-tuned with a percentage of labels with the backbone frozen to compare how well the respective models evaluate against a supervised ResNet18 model. For linear evaluation, early stopping was enabled when training failed to decrease in loss for 10 epochs.

3.1 Experimental Setup

In assessing the efficacy of the representations generated by various SSL frameworks, linear evaluation is used to assess the quality of the learned representations. The linear evaluation method involves training a supervised linear classifier on the SSL pre-trained models, with the model weights kept constant. The classification score derived from this process provides insight into the discriminative capacity of the pre-trained representations and serves as an indirect measure of the model's performance in subsequent tasks [28].

3.2 Dataset

Labeled multi-band SAS data is hard to come by and quite limited. Due to the high resolution nature of SAS data, it is often too large for modern GPUs requiring the imagery to be broken up into tiles/snippets/chips. To generate a dataset that consists of snippets, a Reed-Xiaoli [29] anomaly detector is used to to detect potential objects of interest by extracting snippets from high-resolution SAS imagery. The low and high frequency, call them LF and HF respectively, snippets are first resized to 224x224 each and stacked forming a 2x224x224 multi-band SAS image. Previous works have applied the multi-band approach to success [30–32]. Different beamformers have been used to generate the SAS imagery providing semantically the same scenes to the human eye yet statistically different.

The collective snippets make up the four datasets used in experiments: pre-train, train, validation, and test. For simplicity, the labeled datasets(train, validation, and test) used in linear probing experiments are balanced for positive and negative instances.

3.3 SSL Models and Hyperparameters

This work leverages two different types of SSL model architectures: MoCov2 [1] and BYOL [2]. The two SSL methods have been categorized as contrastive and non-contrastive.

- **MoCov2:** displays strength in learning meaningful representations by contrasting positive and negative samples where distinguishing between different classes is important. However it's contrastive strength, it requires careful selection of negative samples and the size of the queue can significantly affect the performance.
- **BYOL:** a popular non-contrastive SSL method as it avoids the need for negative samples which can simplify the training process and reduces computational requirements that contrastive loss functions require such as large batch sizes. The non-contrastive method does come at cost where distinguishing between different classes is crucial.

Hyperparameter	MoCov2	BYOL
Backbone	ResNet18	ResNet18
Channels	2	2
Epochs	100	100
Optimizer	AdamW	AdamW
Scheduler	Cos Anneal	Cos Anneal
Loss	NTXent	Neg Cos
Learning Rate	0.003	0.003
Weight Decay	0.001	0.001
Batch Size	768	512

Table 1: Two different types of SSL models, MoCov2 and BYOL, with similar hyperparameters.

3.4 Data Augmentations

Data augmentation techniques, which generate diverse and challenging examples, are heavily relied upon in SSL. Ensuring that the augmentations are diverse and cover a wide range of transformations can help prevent overfitting thus a moderate amount of augmentations are lightly applied to drive the feature learning process during pre-training as shown in Figure 2. Speckle noise is artificially introduced into the SAS image that multiplies a constant noise factor across the imagery. During training, only horizontal flip augmentations were applied.

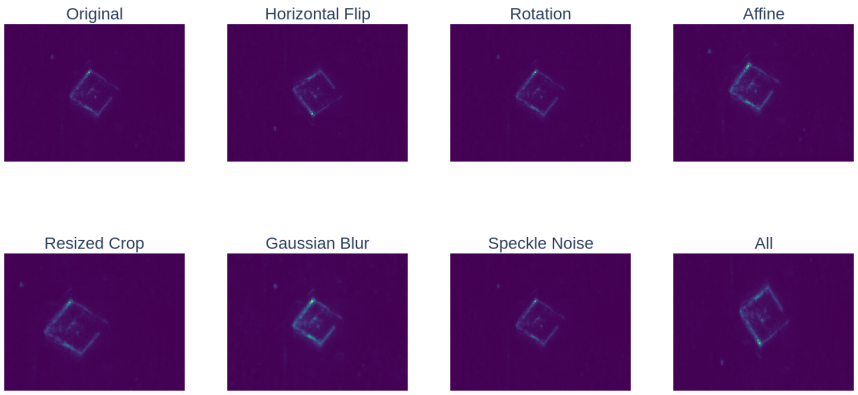

Figure 2: Visualization of the SAS data augmentation pipeline for a box-like object during pre-training to create a drastic contrastive image.

3.5 Performance Metrics

In the context of image classification with SAS, the evaluation metrics and benchmarks used for binary image classification tasks need to effectively measure the ability of the model to accurately distinguish between objects and objects not of interest (typically representing seafloor clutter or other underwater objects). The following performance metrics were used to evaluate the models:

- **Contrastive loss:** During pre-training, the contrastive loss is tracked on a validation dataset providing insight on how well the model is learning to distinguish between similar and dissimilar samples.
- **Recall (Sensitivity or True Positive Rate):** This measures the proportion of actual positives (objects of interest) that were identified correctly. A high recall is crucial in image classification because failing to identify an object (false negative), could be disastrous in contested military waters.
- **Precision (Positive Predictive Value):** This measures the proportion of positive identifications (identified objects) that were actually correct. A high precision means a low false positive rate, which is desirable in image classification tasks to avoid wasting resources on false detections.
- **Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC):** This metric provides a comprehensive measure of performance across all possible classification thresholds, summarizing the trade-off between the true positive rate and false positive rate.
- **Accuracy:** This is the simplest metric, representing the proportion of total predictions that were correct. However, accuracy can be misleading if the classes are imbalanced (e.g., if objects are much less common than distractor objects).

4 RESULTS

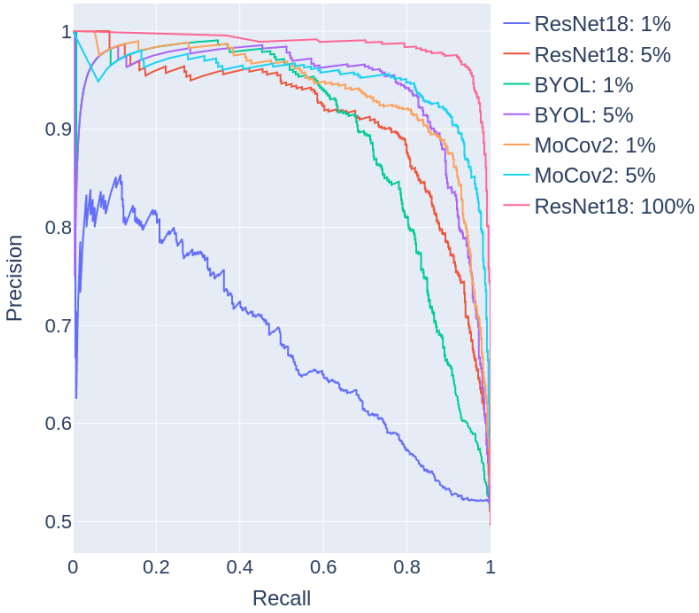

Figure 3: Precision-Recall curves demonstrate SSL model trade-offs for varying labeled scenarios.

The integration of SSL into SAS significantly enhances the performance of the SSL models, specifically MoCov2 and BYOL, when only 1% and 5% of the labels are utilized during training. However, when compared to the ResNet18 model, which had access to 100% of the labels, the SSL models fell short, as shown in Figures 3, 4, and 5. The SSL algorithms were able to effectively extract high-level features from the SAS data, resulting in enhanced performance in downstream tasks for limited label scenarios.

5 DISCUSSION

The results suggest that SSL can be effectively applied to SAS, similar to its successful application in SAR and other computer vision tasks. The improved performance in image classification tasks indicates the potential of SSL in enhancing SAS target recognition for low labeled regimes. However, when abundant data labels exist, supervised learning outperforms in all aspects.

In order to better understand the feature representations learned by the models, t-SNE [33], a popular technique for visualizing high-dimensional data was deployed. Figure 6 shows the t-SNE visualizations of the feature representations learned by MoCov2 and BYOL.

As can be seen from the visualizations, both the SSL models and the supervised model have learned to cluster the sonar images in a meaningful way, with images of the same class clustering together. This suggests that the models have learned to extract features that are relevant for the task of sonar object classification. More compact and well-separated clusters indicate the SSL models learned robust and discriminative features, which could potentially lead to better performance in downstream tasks other than image classification such as object detection, segmentation, and change detection.

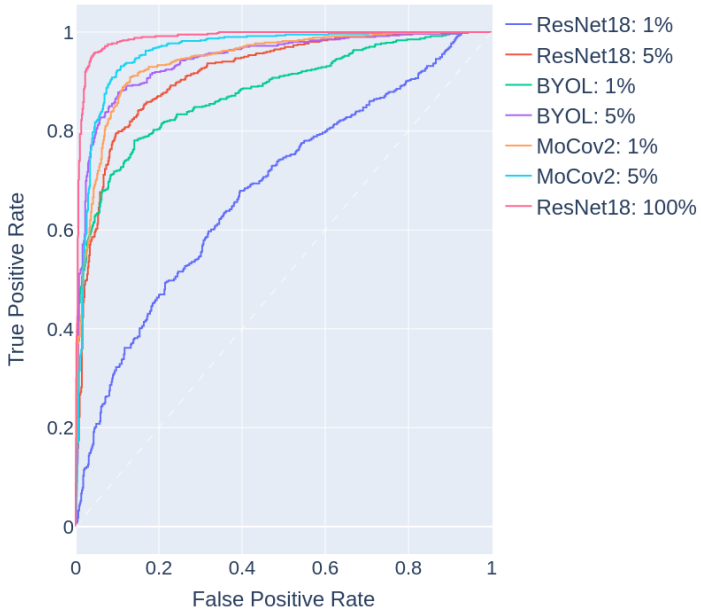

Figure 4: ROC curves show SSL model trade-offs for varying labeled scenarios.

5.1 Implication of Results

Based on the findings, the application of SSL to SAS significantly improves the performance of target recognition tasks for low labeled regimes. This has several important implications.

Firstly, it suggests that SSL can effectively leverage the abundance of unlabeled SAS data, which has traditionally been a challenge in this field. This could potentially revolutionize the way we process and analyze SAS data, leading to more efficient and cost-effective methods.

Secondly, the improved performance in downstream tasks such as image classification indicates that SSL can enhance the practical utility of SAS in various applications, such as underwater exploration, marine archaeology, and other naval applications.

Finally, the results contribute to the growing body of evidence supporting the use of SSL in remote sensing and could stimulate further research in this area.

6 CONCLUSION

The potential of self-supervised learning to improve the classification of SAS images is underscored in this study. Given their success in various computer vision tasks, future research could explore the use of Vision Transformers (ViTs) as backbones for SSL with SAS data.

Additionally, a multi-modal SSL approach that leverages all available data collected by autonomous underwater vehicles, such as bathymetric data or other sonar modalities, could potentially provide richer representations and improve performance.

While the application of SSL to SAS tasks is promising, it is still in its infancy. Further exploration could significantly advance automated underwater computer vision tasks.

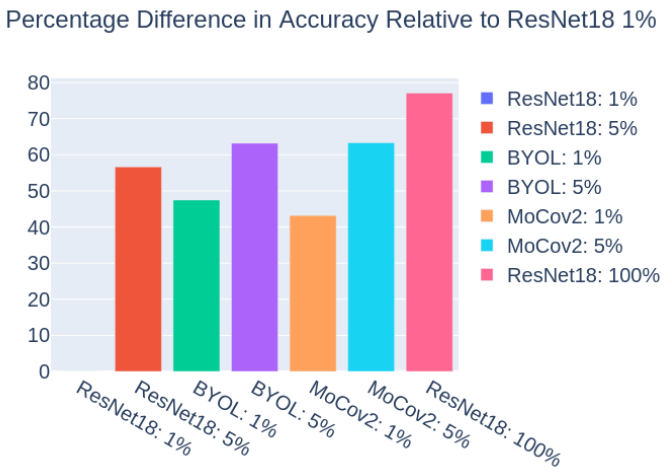

Figure 5: Relative accuracy to ResNet18 1% increases for all other models during evaluation.

Figure 6: MoCov2 and BYOL t-SNE representations show how well each model clusters SAS images.

7 REFERENCES

[1] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *ArXiv*, abs/2003.04297, 2020.

[2] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Najd Alosaimi, Haikel Salem Alhichri, Yakoub Bazi, Belgacem Ben Youssef, and Naif A. Alajlan. Self-supervised learning for remote sensing scene classification under the few shot scenario. *Scientific Reports*, 13, 2023.

[5] Paul Berg, Minh-Tan Pham, and Nicolas Courty. Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives. *Remote. Sens.*, 14:3995, 2022.

[6] Chao Tao, Ji Qi, Mingning Guo, Qing Zhu, and Haifeng Li. Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. *ArXiv*, abs/2211.08129, 2022.

[7] Sachith Seneviratne, Jasper S. Wijnands, Kerry A. Nice, Haifeng Zhao, Branislava Godic, Suzanne Mavoa, Rajith Vidanaarachchi, Mark Stevenson, Leandro M. T. Garcia, Ruth F. Hunter, and Jason Thompson. Urban feature analysis from aerial remote sensing imagery using self-supervised and semi-supervised computer vision. *ArXiv*, abs/2208.08047, 2022.

[8] Heechul Jung, Yoonju Oh, Seongho Jeong, Chaehyeon Lee, and Taegyun Jeon. Contrastive self-supervised learning with smoothed representation for remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[9] Vladimir Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1182–1191, 2021.

[10] Xiaomin Li, D. Shi, Xiaolei Diao, and Hao Xu. Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

[11] Yi Wang, Conrad M. Albrecht, and Xiaoxiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 139–142, 2022.

[12] Chaitanya Patel, Shashank Sharma, and Varun Gulshan. Evaluating self and semi-supervised methods for remote sensing segmentation tasks. *ArXiv*, abs/2111.10079, 2021.

[13] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. *ArXiv*, abs/2211.07044, 2022.

[14] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, M. Burke, D. Lobell, and Stefano Ermon. Geography-aware self-supervised learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10161–10170, 2020.

[15] Alan Preciado-Grijalva, Bilal Wehbe, Miguel Bande Firvida, and Matias Valdenegro-Toro. Self-supervised learning for sonar image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1499–1508, 2022.

[16] Johnny Chen and Jason E Summers. Deep convolutional neural networks for semi-supervised learning from synthetic aperture sonar (sas) images. In *Proceedings of Meetings on Acoustics 173EAA*, volume 30, page 055018. Acoustical Society of America, 2017.

[17] Angeliki Xenaki, Bart Gips, and Yan Pailhas. Unsupervised learning of platform motion in synthetic aperture sonar. *The Journal of the Acoustical Society of America*, 151(2):1104–1114, 2022.

[18] Yung-Chen Sun, Isaac D Gerg, and Vishal Monga. Iterative, deep, and unsupervised synthetic aperture sonar image segmentation. In *OCEANS 2021: San Diego–Porto*, pages 1–5. IEEE, 2021.

[19] Paul Berg, Minh-Tan Pham, and Nicolas Courty. Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives. *Remote Sensing*, 2022.

[20] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *Ieee Geoscience and Remote Sensing Magazine*, 2022.

[21] Zachary D. Calhoun, Saad Lahrichi, Simiao Ren, Jordan M. Malof, and Kyle Bradbury. Self-supervised encoders are better transfer learners in remote sensing applications. *Remote Sensing*, 2022.

[22] Jules Bourcier, Gohar Dashyan, Jocelyn Chanussot, and Karteek Alahari. Evaluating the label efficiency of contrastive self-supervised learning for multi-resolution satellite imagery. In *Image and Signal Processing for Remote Sensing XXVIII*, volume 12267, pages 152–161. SPIE, 2022.

[23] Huihui Dong, Wenping Ma, Yue Wu, Jun Zhang, and Licheng Jiao. Self-supervised representation learning for remote sensing image change detection based on temporal prediction. *Remote Sensing*, 2020.

[24] Linus Scheibenreif, Michael Mommert, and Damian Borth. Contrastive self-supervised data fusion for satellite imagery. *Isprs Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences*, 2022.

[25] Jue Wang, Yanfei Zhong, and Liangpei Zhang. Change detection based on supervised contrastive learning for high-resolution remote sensing imagery. *Ieee Transactions on Geoscience and Remote Sensing*, 2023.

[26] Zhicheng Zhao, Ze Luo, Jian Li, Can Chen, and Yingchao Piao. When self-supervised learning meets scene classification: Remote sensing scene classification based on a multitask learning framework. *Remote Sensing*, 2020.

[27] Antonio Montanaro, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Semi-supervised learning for joint sar and multispectral land cover classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[28] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1920–1929, 2019.

[29] Irving S Reed and Xiaoli Yu. Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. *IEEE transactions on acoustics, speech, and signal processing*, 38(10):1760–1770, 1990.

[30] David P. Williams. On the utility of multiple sonar imaging bands for underwater object recognition. In *OCEANS 2022, Hampton Roads*, pages 1–6, 2022.

[31] M Emigh, B Marchand, M Cook, and J Prater. Supervised deep learning classification for multiband synthetic aperture sonar. In *Proceedings of the 4th International Conference on Synthetic Aperture Sonar and Synthetic Aperture Radar*, volume 40, pages 140–147, 2018.

[32] Isaac D. Gerg. Multiband sas imagery. *ArXiv*, abs/1808.02792, 2018.

[33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.